REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*: 30-11-2002

2. REPORT TYPE: Final Technical Report

3. DATES COVERED *(From – To)*: 1-6-2001

4. TITLE AND SUBTITLE: Eye Tracking While Answering Questions in Electronic Multimedia Environments

5a. CONTRACT NUMBER: 01-06-061

5b. GRANT NUMBER: N00014-01-1-0917

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHOR(S): Graesser, Arthur C.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Psychology
202 Psychology Building
The University of Memphis
Memphis, TN 38152-3230

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Ballston Centre Tower One
800 North Quincy Street
Arlington, VA 22217-5660

10. SPONSORING/MONITOR'S ACRONYM(S): ONR

11. SPONSORING/MONITORING REPORT NUMBER

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This research project collected eye tracking data while adults answered questions, asked questions, and interacted with electronic media. We tested computational models of question answering (QUEST), question comprehension difficulty (QUAID), and question asking (PREG) in four contexts: (1) answering questions on government questionnaires, (2) comprehending illustrated texts on mechanical and electronic devices, (3) reading Navy recruiting material on the web, and (4) engaging in collaborative dialog with an intelligent computer tutor (AutoTutor) that teaches students computer literacy. Five eye tracking experiments were conducted in these empirical tests of the models. Measures of eye tracking in these tasks were correlated with measures of individual differences in some of these experiments. This research has direct applications to survey research methodology, recruiting, training, and selection & classification.

15. SUBJECT TERMS
Question asking, question answering, cognitive modeling, eye tracking, survey methodology

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON / 19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	U	6	Arthur C. Graesser / (901) 678-2742

FINAL TECHNICAL REPORT

GRANT #: N00014-01—1-0917

PRINCIPAL INVESTIGATOR: Arthur C. Graesser (e-mail: a-graesser@memphis.edu)

INSTITUTION: University of Memphis

GRANT TITLE: Eye Tracking While Answering Questions in Electronic Multimedia Environments

AWARD PERIOD: 1 June 2001 - 30 September 2002

OBJECTIVE: To collect eye tracking data in order to test predictions of cognitive computational models of question comprehension (QUAID), question asking (PREG), and question answering (QUEST, AutoTutor) when adults interact with electronic multimedia environments.

APPROACH: The acquisition of eye tracking equipment (funded primarily by ONR) has given us the opportunity to investigate cognitive mechanisms in different cognitive tasks. In particular, we measured eye fixation durations and sequences of eye movements while adults comprehend, ask and answer questions during the course of interacting with several complex computer artifacts: computerized surveys and questionnaires, multimedia with illustrated texts, Navy recruiting material on the web, and intelligent tutoring systems with animated conversational agents. We developed most of these computer artifacts with question facilities on other funded grants (i.e., QUEST, QUAID, PREG, and AutoTutor). This research focused on testing cognitive mechanisms by collecting eye tracking data while the adults use these artifacts.

ACCOMPLISHMENTS: Questionnaires, information retrieval systems, computer-based training systems, and computer-human interfaces need to be improved to fit the constraints of human cognition. The comprehension, asking and answering of questions are core cognitive components that drive thought and behavior in virtually any task, so it is important to understand the underlying cognitive mechanisms and to build artifacts around them. The long-term goals of this research are to help us improve these artifacts and to test cognitive models. We conducted 5 experiments that collected eye tracking data on four tasks.

Task 1: Comprehension Problems with Questions on Government Questionnaires. When respondents do not understand the meaning of a survey question, it is very unlikely that they will supply valid and reliable answers. Survey methodologists would therefore benefit from computer tools and other analytical schemes that help them identify problems with questions with respect to comprehension difficulty. We developed a web facility (called QUAID, www.psyc.memphis.edu/quaid.html) that assists survey methodologists in improving the wording, syntax, and semantics of questions (Graesser, Wiemer-Hastings, Kreuz, Wiemer-Hastings, & Marques, 2000). QUAID stands for "Question Understanding Aid." The survey methodologist enters a question on a questionnaire, along with context information and answer alternatives that accompany the question. QUAID quickly returns a list of five classes of potential problems with the question: (1) unfamiliar technical terms, (2) vague or imprecise relative terms, (3) vague or ambiguous noun-phrases, (4) complex syntax, and (5) working memory overload. Recent advances in cognitive science, computational linguistics and discourse processing have reached the point where it possible to analyze language and meaning at these various levels (Jurafsky & Martin, 2000). Graesser, Wiemer-Hastings et al. (2000) reported how well QUAID diagnoses the five categories of problems with questions in a corpus of questions provided by the US Census Bureau. Experts in language and cognition rated each question as to whether it had the five problems; the expert ratings were compared with QUAID output. Analyses of hit rates, false alarm rates, and discrimination scores confirmed that there is a significant correspondence between QUAID and the judgments of experts.

On the present grant, two eye tracking experiments were conducted that collected data from respondents as they answered questions. If a question is problematic, this should be reflected in the patterns of eye movements and the amount of time that respondents focus on individual words. In Experiment 1, 24 college students answered 69 survey questions presented on a computer screen, one at a time. These

questions were sampled from 11 different surveys provided by the US Census Bureau. Most of the questions (72%) had comprehension problems according to QUAID and expert survey methodologists. We collected data from 32 students in Experiment 2. Whereas Experiment 1 randomly presented questions from different surveys, Experiment 2 had a question presentation order that followed the order of questions on the surveys provided by the US Census Bureau. We also are administering the Wechsler Abbreviated Scale of Intelligence (WASI) and other tests of individual differences. Participants were classified into high versus low verbal ability and other measures of individual differences. However, the measures of individual differences rarely had a significant impact on eye tracking behavior so they are not included in this report.

One of our discoveries challenges some of the conventional models of reading outside of the arena of survey questions. Such models assume that more time is spent reading verbal material that is more complex. In contrast, we discovered that the eye tracking data was robustly explained by the processes of exiting a question (as in the case of generating an answer before finishing the reading of the question). Obviously, the validity of answers to questions is seriously compromised if readers have an early exit from question comprehension. A substantial proportion of the fixation time variance is explained by the line number and the line position. For example, we conducted an analysis on the content words (i.e., nouns, main verbs, adjectives) but not function words (e.g., prepositions, articles). The mean fixation times per content word on lines 1-3 were 333, 275, versus 162 milliseconds for the left, middle, and right positions, whereas the corresponding means were 335, 291, and 140 for lines 4-6 and were 201, 140, and 126 for lines 7-10. Therefore, 2-3 times as much time is allocated to content words in the upper left quadrant of the display than the lower right quadrant. These layout features masked the effects of the difficulty of question comprehension. Two approaches were pursued in order to have a more sensitive analysis of the impact of question comprehension problems on eye tracking behavior. First, we conducted analyses that statistically controlled for line number and line position. Second, we inspected the entire frequency distribution of the processing time on individual words (i.e., total fixation times on a word, first fixation times, likelihood that a word is skipped). This second approach provided the most informative results.

We segregated the frequency distribution of fixation times into partitions of 50 milliseconds and computed the proportion of fixations in each partition. The partition of 0 to 49 milliseconds was noteworthy because words in this increment were defined as skipped. The content words were sorted into Difficult versus Easy from the standpoint of each of the five classes of problems with questions (1 through 5 above). An analysis of variance (ANOVA) was performed on the proportion scores in a factorial design that crossed Problem (Difficult vs. Easy) with time Partition (with anywhere from 20 to 40 partitions in an analysis). There was a statistically significant Problem X Partition interaction for 4 out of 5 classes of problems identified by QUAID. This result supports the claim that QUAID's identification of problematic words and questions was indeed manifested in the patterns of eye tracking behavior. For example, consider the total fixation times of unfamiliar technical terms versus other content words, which had a significant Problem x Partition interaction, $F(18, 504) = 2.14$, $p < .05$. The likelihood of skipping a word was significantly lower for unfamiliar technical terms (.17) than for easy content words (.35). First fixation times had a mode (i.e., partition with the highest frequency) of 151-200 milliseconds for unfamiliar technical terms, which is longer than the mode of 101-150 milliseconds for easy content words. Therefore, the words QUAID identified as problematic did significantly affect eye tracking behavior, which is a testimony to its validity. Technical terms were not skipped as often, and were processed for a longer duration. The other classes of problems also were reflected in eye tracking behavior, but it is beyond the scope of this report to discuss them.

Task 2: Illustrated texts on mechanical and electronic devices. College students read illustrated texts on mechanical and electronic devices, such as cylinder locks, toasters, and dishwashers. One research goal was to investigate their patterns of eye movements during normal comprehension of the illustrated texts. We collected eye tracking data from 40 students as they normally read 5 different illustrated texts (3 minutes per text). We examined how much time was devoted to fixating on the text, the illustrations, the word labels in the diagrams, versus the directional arrows that depict motion. We discovered that the students allocate 6 times as much time to the text than to the other components of the display (illustrations, labels, and directional arrows). The fact that the text reigns supreme in commanding the reader's attention replicates results reported by Hegarty and Just (1993), but our study is the first that dissects attention to arrows, labels, versus illustrations. We also correlated the allocation of eye fixations with measures of individual

differences and device comprehension scores. We discovered that the depth of device comprehension (as measured by an objective multiple choice test) was positively correlated with the attention to the illustrations and arrows (i.e., pictorial information, correlations of .45 to .55) but not to the text and word labels (i.e., verbal information, correlations near zero). Moreover, we found that deep comprehenders spent more time integrating information from the text and picture.

Nevertheless, we believe that deep comprehension of a device is rarely manifested when adults read text for no particular purpose. It is not until a device breaks down, for example, that they realize how little they know about the inner workings. From the standpoint of cognitive theory (Kieras & Bovair, 1984), everyday knowledge about devices primarily consists of perceptual knowledge of the outer surfaces (i.e., what the device looks like) and procedures for starting, monitoring, and terminating the device (i.e., how to use it), but not the mental models of the device mechanisms (how it works). The need for mental models of the device mechanisms exists when devices break down, but rarely when the devices are properly functioning. Deeper mental models are needed to diagnose malfunctions and discovering methods of repairing the devices. Our ONR research revealed that the depth of a person's comprehension of a device is manifested by the questions they ask when the device breaks down (Graesser & Olde, 2001). Deep comprehenders ask good questions that inquire about the likely faults of the breakdowns.

A cognitive model of question asking (called PREG, Otero & Graesser, 2001) assumes that questions emerge when there is cognitive disequilibrium, as in the case of contradictions, obstacles to goals, anomalies, salient contrasts, and major gaps in knowledge. Eye tracking data were collected to investigate the process of question asking when individuals are confronted with cognitive disequilibrium. College students (N=40) read illustrated texts about everyday devices (e.g., a cylinder lock) and then were given a breakdown scenario (e.g., the key turns but the bolt does not move). They were instructed to ask questions when given the breakdown scenario and the eye tracker recorded their eye fixations. As predicted by the PREG model, college students who understood the devices at a deep level also asked better questions (r = .45) and their eyes fixated for more time on the device components that would explain the malfunction (r = .52). The fixation on plausible faults primarily occurred 1-3 seconds immediately before the question was articulated, which suggests that disequilibrium and fault detection are initiated before question articulation. These results not only supported the predictions of the PREG model of question asking, but supported the conclusion that a quick litmus test of whether a person deeply understands a device is whether their eyes tend to focus on likely faults when presented a breakdown scenario.

Task 3: Reading Navy recruiting material. The Navy has a web site that attempts to recruit young adults into the Navy (www.navy.mil) and to reveal the diverse jobs that are available to sailors (www.navyjobs.com). Eye tracking data were collected while college students perused 12 of these web pages for the purpose of answering questions. College students (N=27) were randomly assigned to one of three conditions with different orienting questions: (1) "What are the educational and financial benefits in joining the Navy?", (2) "What are the requirements for joining the Navy?", versus (3) a no-leading-question control. There were web pages that included information that answered the first question, some pages that addressed the second, some pages that addressed both questions, and some pages that addressed neither question. According to the QUEST model of question answering (Graesser, Gordon, & Brainerd, 1992), the orienting questions were expected to systematically influence the length of time a screen was read and also the fixation durations on particular regions of the screen.

Both of the predictions of QUEST of question answering were supported. There was a significant interaction between the 3 conditions and 4 categories of web pages, i.e., whether there was information that answered the orienting questions, $F(6, 72) = 2.46, p < .05$. Planned comparisons confirmed that pages were read longer if there was relevant information that answered an orientation question than if they had no relevant information, 159 versus 110 seconds; the time reading the pages in the control condition was 103 seconds per page. We identified areas of interest on the web pages that had information that directly answered an orientation question, as opposed to those areas of interest with no relevant information. The initial fixation time on an area of interest (i.e., not including re-fixations) was 209 milliseconds longer for those areas of interest with answer information. These results support QUEST. Comparisons to the control condition also supported two mechanisms: (1) More time is spent processing relevant information that

answers a specific orienting question and (2) less time is spent processing irrelevant information. The next step is to contrast the predictions of QUEST with predictions of a bottom-up perception model that assumes that fixations are captured by pictures, contrasts, headers, and other salient areas of the web pages.

Task 4: Tutoring systems with conversational dialog. Animated conversational agents have facial features synchronized with speech and in some cases appropriate gestures (Johnson, Rickel, & Lester, 2000). The gestures point to important material on display and emphasize ideas. The text-to-speech engines can articulate any text that gets passed to the agent, with timing and intonation that is surprisingly similar to human speech. In essence, the agents provide an anthropomorphic human-computer interface that simulates having a conversation with a human.

In recent years, educational researchers have explored whether these conversational agents can be effectively integrated with learning environments. For example, there is some evidence that these agents have a positive impact on learning or on the learner's perceptions of the learning experience when subject matter is presented with the agents, compared with speech alone or text controls (Atkinson, 2002). Animated agents have recently been integrated with intelligent tutoring systems that hold dialogues with students in natural language. One notable example is AutoTutor (Graesser, Person, Harter, 2001; Graesser, VanLehn, Rose, Jordan, & Harter, 2001), a web based intelligent tutoring system that has shown significant learning gains (effect sizes of .5 to 1.5 standard deviation units). AutoTutor is not merely an information delivery system, but serves as a discourse prosthesis (or collaborative scaffold) that responds appropriately to the student's conversational turn and that assists the student in actively constructing knowledge. For example, AutoTutor gives feedback to the student on what the student types in (positive, neutral, negative), pumps for information ("What else?"), prompts for specific information, gives hints, fills in missing information with assertions, corrects misconceptions, and summarizes answers.

We conducted a study that collected eye tracking data while 12 college students interacted with AutoTutor in a tutorial dialogue on computer literacy. Eye tracking data were collected during a 20-minute time span while the participants first interacted with AutoTutor. We explored how the learners allocated their attention to four windows in the display during the tutoring session: (1) The animated conversational agent (which took up 28% of the computer screen), (2) the main deep-reasoning question (e.g., why, how, what-if, 13% of the screen), (3) a graphic display of the computer system components that are relevant to the question (46% of the display), and (4) the student's input (a window that shows what the student types in, 13% of the display). The animated agent could conceivably affect eye tracking in a number of different ways. One possibility is that the agent is an "attention magnet" that draws the learner's attention more than would be expected by chance, if eye movements were randomly allocated to regions on the display. Another possibility is that that there is a "novelty effect." That is, the agent might start out being a magnet of attention, but attention on it decreases over time. In order to assess this possibility, we measured the fixation times on the four information sources as a function of four 5-minute time intervals in the 20-minute tutoring session. We computed a metric called an adjusted fixation ratio (AFR), which is 1 if the eyes wander randomly within the computer screen, greater than 1 to the extent that an agent is an attention magnet, and less than 1 to the extent the agent is ignored.

Our first analysis measured the amount of time the eyes fixated on the four windows of the display as a function of the four 5-minute time intervals. There were statistically significant differences in the percentage of time allocated to the four information sources, with means of 54, 5, 32, and 9%, for the agent, main question, graphic display, and student input regions, respectively, $F(3,33) = 31.11$, $p < .01$, but there was no main effect or interaction with time interval. The AFR scores were 1.81, .43, .78, and .68 for the corresponding four information sources. These results robustly support the attention magnet hypothesis but not the novelty effect hypothesis. The animated agent is clearly a strong magnet of attention, and this effect does not decrease over time.

A secondary question explored eye fixations on different parts of the talking head. The AFR scores showed the following ordering: nose (10.33), eyes (10.28), cheeks (8.96), forehead (5.61), mouth (2.38), background (1.81), and shoulders (1.37). The finding that there is an emphasis on the eyes was quite expected because emotions and attention are conveyed through the eyes (Massaro & Cohen, 1995). The

finding that comparatively little attention was allocated to the mouth was not expected, because emotions are conveyed in mouth shapes and sometimes ambiguities in speech can be resolved by viewing the mouth. Apparently the speech quality was adequate so that redundancy in cues was not helpful. The comparatively high allocation of attention to the nose and cheeks was entirely unexpected. Perhaps the nose attracts attention because it is visually salient and is at the center of the face. We do not have a good explanation for attention on the cheeks, but they are midpoint between the eyes and the corners of the mouth. Additional research is needed to explore the generality of these results and to dissect potential explanation.

<u>CONCLUSIONS</u>: Eye tracking data supported a number of predictions of our cognitive models of question comprehension (QUAID), question asking (PREG), and question answering (QUEST, AutoTutor) in the context of complex multimedia environments. We also discovered some trends that pertain to properties of artifacts and multimedia that are outside of the scope of these models. For example, readers exit the reading of a question before all of the words are processed and an answer is generated. The layout of a question is nontrivial because more time is spent on words in the upper left quadrant than the lower right. Good comprehenders tend to allocate more time to the illustrations and directional arrows of illustrated texts on devices, compared with the text and word labels. Talking heads are magnets of the students' attention as they interact with AutoTutor and this trend persists over time.

<u>SIGNIFICANCE</u>: This research addresses many of the missions that were articulated in ***Sailor 21: A Research Vision to Attract, Retain, and Utilize the 21st Century Sailor***: recruitment, training, survey methodology, and personnel classification. This research advances cognitive models of question comprehension, question asking, question answering, human-computer interaction, and the use of complex multimedia. Questions are at the heart of virtually any task an adult performs when using computer technologies. It could be argued that any given task can be decomposed into a set of questions that a sailor asks and answers. When a sailor encounters a device that malfunctions, the relevant questions are "What's wrong?" and ""How can it be fixed?". When an officer reads a technical document, the relevant questions are "Why is this important?" and "What should I do about it, if anything?". When a young adult reads Navy recruiting material, the relevant questions are "What's interesting?", "Do I want to join?", and "What are the perks?". The mechanisms that trigger question asking, exploration patterns, and question answering strategies need to be understood in order to design the messages and technological artifacts effectively.

References

Atkinson, R.K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology, 94*, 416-427.

Graesser, A. C., Gordon, S. E., & Brainerd, L. E. (1992). QUEST: A model of question answering. *Computers and Mathematics with Applications, 23*, 733-745.

Graesser, A.C., & Olde, B.A. (in press). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*.

Graesser, A.C., Person, N., Harter, D., & TRG (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education, 12*, 257-279.

Graesser, A.C., VanLehn, K., Rose, C., Jordan, P., and Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22, 39-51.

Graesser, A.C., Wiemer-Hastings, K., Kreuz, R., & Wiemer-Hastings, P. , & Marques, K. (2000). QUAID: A questionnaire evaluation aid for survey methodologists. *Behavior Research Methods, Instruments, and Computers, 32*, 254-262.

Hegarty, M., & Just, M.A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32*, 717-742.

Johnson, W. L., & Rickel, J. W., & Lester, J.C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11, 47-78.

Jurafsky, D., & Martin, J.E. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, NJ: Prentice Hall.

Kieras, D.E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science, 8*, 255-274.
Massaro, D.W., & Cohen, M.M. (1995). Perceiving talking faces. *Psychological Science, 4*, 104-109.
Otero, J., & Graesser, A.C. (2001). PREG: Elements of a model of question asking. *Cognition & Instruction, 19*, 143-175.

PATENT INFORMATION: No patents have been files.

AWARD INFORMATION: No awards were given.

REFEREED PUBLICATIONS:

1. Otero, J., & Graesser, A.C. (2001). PREG: Elements of a model of question asking. Cognition & Instruction, 19, 143-175.

2. Graesser, A.C., & Olde, B.A. (in press). How does one know whether a person understands a device? The quality of questions the person asks when the device breaks down. Journal of Educational Psychology.

3. Graesser, A.C., Olde, B., & Lu, S. (2001). Question-driven explanatory reasoning about devices that malfunction. In T. Filjak (Ed.), Proceedings of the 36th International Applied Military Psychology Symposium (pp. 114-119). Zagreb, Croatia: Ministry of Defense of the Republic of Croatia.

4. Lu, S., Olde, B., Cooper, E., & Graesser, A.C. (2001). Individual differences in reasoning about broken devices: An eye tracking study. In J.D. Moore and K. Stenning (Eds.), Proceedings of the 23rd Annual Conference of the Cognitive Science Society (pp. 582-587). Mahwah, NJ: Erlbaum.

BOOK CHAPTERS, SUBMISSIONS, ABSTRACTS AND OTHER PUBLICATIONS:

1. Graesser, A.C., Karnavat, A.B., Daniel, F.K., Cooper, E., Whitten, S.N., & Louwerse, M. (2001). A computer tool to improve questionnaire design. In Statistical Policy Working Paper 33, Federal Committee on Statistical Methodology (pp. 36-48). Washington, DC: Bureau of Labor Statistics.

2. Graesser, A.C., Leon, J.A., & Otero, J.C. (2002). Introduction to the psychology of science text comprehension. In J. Otero, J.A. Leon, & A.C. Graesser (Eds). The psychology of science text comprehension (pp. 1-15). Mahwah, NJ: Erlbaum.

3. Graesser, A.C., Olde, B., Pomeroy, V., Whitten, S., Lu, S., & Craig, S. (in press). Inferences and questions in science text comprehension. In J. Otero and M. H. Caldeira (Eds.), Comprension de los libros de texto de ciencias. Madrid: Ed. Paidos.

4. Graesser, A.C., & Wisher, R.A. (2001). Question generation as a learning multiplier in distributed learning environments. United States Army Research Institute for the Behavioral and Social Sciences, Technical Report 1121.

5. Lu, S., & Graesser, A.C. (2001). Does an illustration help to comprehend scientific texts? Proceedings of the Third International Conference of Cognitive Science (Pp. 643). Beijing, China: Press of the University of Science and Technology of China.

6. Olde, B.A., Cooper, E., Lu, S., Whitten, S., & Graesser, A.C. (submitted). Do you look before you speak? Eye tracking and question asking when devices break down. Journal of Experimental Psychology: Learning, Memory, and Cognition.